ADHIRAJ ROKA

STAT- 8446 SURVIVAL DATA ANALYSIS PROJECT

RECURRENCE TIME TO DRUG USE

Contents:

1. INTRODUCTION

The data set used for this project is the subset of data from the University of Massachusetts AIDS Research Unit (UMARU) Impact Study (UIS) taken from the Applied Survival Analysis book of Hosmer, D.W. and Lemeshow, S. and May, S.

The UMARU data set has 628 observations and 12 variables.

The variables in the data set are as follows:

VARIABLE	DESCRIPTION	TYPE	VALUES/CODE
id	Participant id code	Numeric id	1- 628
age	Age at enrollment	continuous	Years
beck	Beck of Depression Score	continuous	0.00-54.00 units
hercoc	Heroin/Cocaine Use During 3 Months Prior to Admission	categorical	1 = Heroin & Cocaine 2 = Heroin Only 3 = Cocaine Only 4 = Neither Heroin nor Cocaine
ivhx	IV Drug Use History at Admission	categorical	1= Never 2= Previous 3= Recent
ndrugtx	Number of Prior Drug Treatments	continuous	0-40 units
race	Participant's Race	categorical	0 = White 1 = Others
treat	Treatment Randomization Assignment	categorical	0 = Short 1 = Long
sit	Treatment Site	categorical	0 = A 1 = B
los	Length of Treatment (measured from admission)	continuous	days
time	Time to return to Drug Use (measured from admission)	continuous	days
censor	Returned to Drug Use	categorical	1 = Returned to Drug Use 0 = Otherwise

- Before looking at the survival times, I wanted to explore to see if the dataset has any missing values or not. The following output shows the missing variables:

Glimpse of the dataset

Obs	id	age	beck	hercoc	ivhx	ndrugtx	race	treat	sit	los	time	censor
1	1	39	9	4	3	1	0	1	0	123	188	1
2	2	33	34	4	2	8	0	1	0	25	26	1
3	3	33	10	2	3	3	0	1	0	7	207	1
4	4	32	20	4	3	1	0	0	0	66	144	1
5	5	24	5	2	1	5	1	1	0	173	551	0
6	6	30	32.55	3	3	1	0	1	0	16	32	1
7	7	39	19	4	3	34	0	1	0	179	459	1
8	8	27	10	4	3	2	0	1	0	21	22	1
9	9	40	29	2	3	3	0	1	0	176	210	1
10	10	36	25	2	3	7	0	1	0	124	184	1

Checking for missing values

The FREQ Procedure

hercoc

hercoc	Frequency	Percent	Cumulative Frequency	Cumulative Percent
.	18	2.87	18	2.87
1	111	17.68	129	20.54
2	114	18.15	243	38.69
3	178	28.34	421	67.04
4	207	32.96	628	100.00

ivhx

ivhx	Frequency	Percent	Cumulative Frequency	Cumulative Percent
.	18	2.87	18	2.87
1	233	37.10	251	39.97
2	115	18.31	366	58.28
3	262	41.72	628	100.00

race

race	Frequency	Percent	Cumulative Frequency	Cumulative Percent
.	6	0.96	6	0.96
0	467	74.36	473	75.32
1	155	24.68	628	100.00

The MEANS Procedure

Variable	Label	N Miss	N
id	id	0	628
treat	treat	0	628
sit	sit	0	628
los	los	0	628
time	time	0	628
censor	censor	0	628
age		5	623
beck		33	595
ndrugtx		17	611

- The red highlighted parts show the missing frequencies.

- We can see from above that some of the variables in the dataset have missing values. Since the number of missing values is not very high and there is no clear reason to have them in the following analysis, I decided to drop all the rows having missing values instead of imputing them for the sake of computation. After dropping these rows, the total number of observations in the dataset are now 575.

2. OBJECTIVES

The objective of the analysis is to build a model to show the return to drug use and look at how the different variables in the model are related to recurrence to drug use. Specifically, those shown below.

The research questions are as follows:

i. To compare the treatment programs of different planned durations designed to reduce drug use.

ii. Test whether the age of the participants and the treatment site for different treatment paired respectively act as confounders or effect modifiers.

iii. Build the best model for modeling survival, i.e., return time to drug use using a cox proportional hazards model.

3. RESULTS

i.

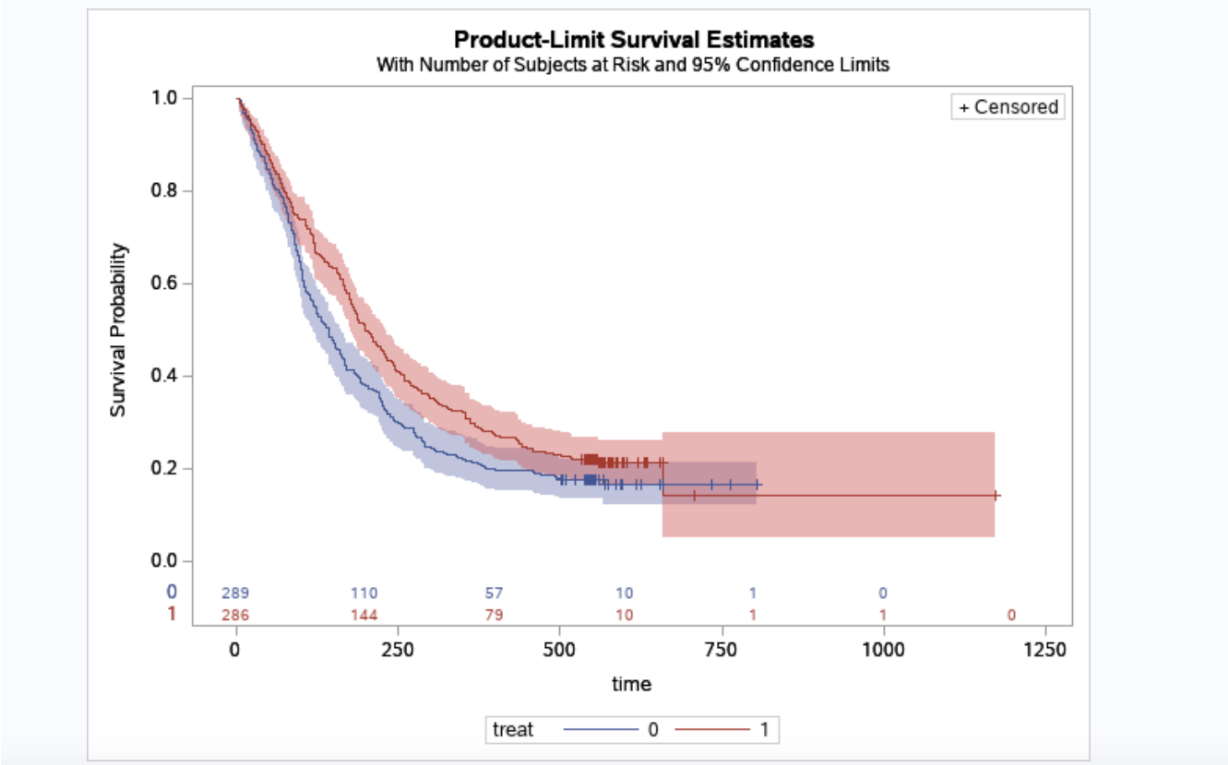
We first look at the time to return to drug use by treatment which is the treat variable in the dataset. The treat variable is a dichotomous variable with 0 for short treatment assigned and 1 for long treatment assigned.

The following output is from the lifetest procedure in SAS:

Test of Equality over Strata			
Test	Chi-Square	DF	Pr > Chi-Square
Log-Rank	6.5284	1	0.0106
Wilcoxon	9.2763	1	0.0023
-2Log(LR)	7.0201	1	0.0081

Adjustment for Multiple Comparisons for the Logrank Test				
Strata Comparison			p-Values	
treat	treat	Chi-Square	Raw	Tukey-Kramer
0	1	6.5284	0.0106	0.0106

Adjustment for Multiple Comparisons for the Wilcoxon Test				
Strata Comparison			p-Values	
treat	treat	Chi-Square	Raw	Tukey-Kramer
0	1	9.2763	0.0023	0.0023



Summarizing the results from the output we have:

Treat= 0 (short)

The median survival time/return to drug use estimate for short treatment is 142 days with a 95% confidence interval of (119.00, 162.00) days.

The 1-, 2-, 3-, and 4-year survival rates for each treatment along with 95% confidence intervals are recorded:

t(i)/days	$\widehat{S}(t_i)$	95% CI
365	0.21453	(0.16923,0.26348)
730	0.16544	(0.12229, 0.21430)
1095	.	.
1460	.	.

Treat= 1 (long)
The median survival time/return to drug use estimate for long treatment is 201.50 days with a 95% confidence interval of (176.00, 231.00) days.

The 1-, 2-, 3-, and 4-year survival rates for each treatment along with 95% confidence intervals:

t(i)	$\widehat{S}(t_i)$	95% CI
365	0.29720	(0.24529,0.35078)
730	0. 14196	(0.05031, 0.27948)
1095	0.14196	(0.05031, 0.27948)
1460	.	.

Note: the time in the dataset does not extend to more than 3 years, hence some estimates are blank.

- Comparing the two tables, especially up to the first two years, and also looking at the survival curves, the time to recurrence to drug use for long treatment is higher.

Furthermore, I used the LOG-RANK TEST to test the significance of treatment on the return time to drug use:

Ho: $S_0(t) = S_1(t)$ vs. Ha: $S_0(t) \neq S_1(t)$, where treat: 0-short, 1-long

$\chi 2$ test stat= 6.5284, df=1, p-value= 0.0106.
Reject Ho. At 0.05 level, there is enough evidence to claim that the return to drug use times between the short treatment and long treatment assigned patients are different. Additionally, looking at the survival curves from the output above, for the majority of time, especially until the first 2 years, the survival curve for the long treatment is higher.

Also, the median return times for long treatments is higher than the short treatment; hence, *treatment: 1-long* seems to be more effective as participants' time to return to drug use is longer.

---

ii.

After finding the treatment variable to be significant. I wanted to model recurrence time to drug use. I did this using the treatment variable paired with age and treatment site by using a Cox proportional hazard model.

First of all, let's look at the age variable. The average age of participants in the short treatment group is 32.66 years, and the average age of participants in the long treatment group is 32.10 years as seen from the mean output below.

The MEANS Procedure

Analysis Variable : age						
treat	N Obs	N	Mean	Std Dev	Minimum	Maximum
0	289	289	32.6574394	6.0481536	20.0000000	53.0000000
1	286	286	32.1048951	6.3347699	21.0000000	56.0000000

After fitting the Cox PH model with treat and age variable to model recurrence time we get the following output:

Model Fit Statistics

Criterion	Without Covariates	With Covariates
-2 LOG L	4978.748	4968.852
AIC	4978.748	4972.852
SBC	4978.748	4981.132

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	9.8959	2	0.0071
Score	9.8638	2	0.0072
Wald	9.8331	2	0.0073

Type 3 Tests

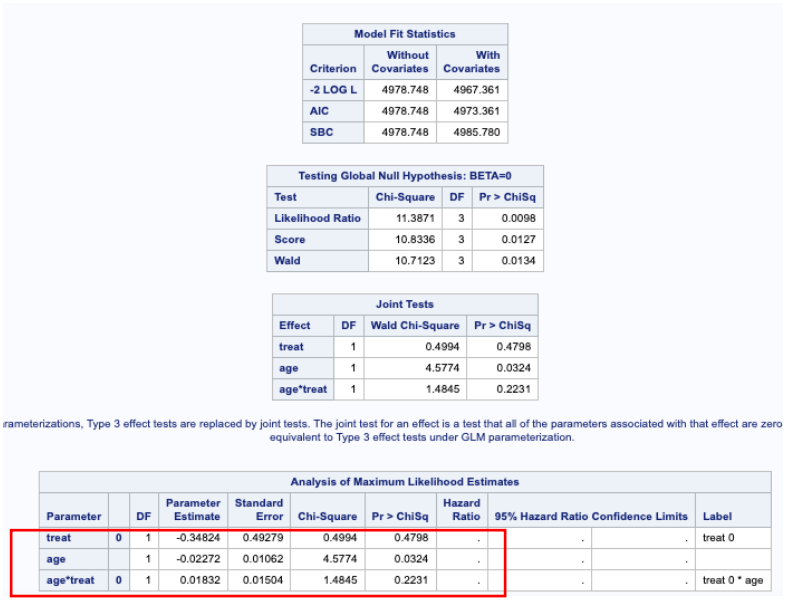
Effect	DF	Wald Chi-Square	Pr > ChiSq
treat	1	6.7369	0.0094
age	1	3.3464	0.0674

Analysis of Maximum Likelihood Estimates

Parameter		DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits		Label
treat	0	1	0.24165	0.09310	6.7369	0.0094	1.273	1.061	1.528	treat 0
age		1	-0.01370	0.00749	3.3464	0.0674	0.986	0.972	1.001	

Looking at the p-values in the table above, we can see that the treat is significant, but the covariate age is not significant at 0.05 level.

Furthermore, I went ahead to test if age was a confounder or an effect modifier in the recurrence model. To do so I also need to fit an interaction model of age x treat and the following is the output:



Model Fit Statistics

Criterion	Without Covariates	With Covariates
-2 LOG L	4978.748	4967.361
AIC	4978.748	4973.361
SBC	4978.748	4985.780

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	11.3871	3	0.0098
Score	10.8336	3	0.0127
Wald	10.7123	3	0.0134

Joint Tests

Effect	DF	Wald Chi-Square	Pr > ChiSq
treat	1	0.4994	0.4798
age	1	4.5774	0.0324
age*treat	1	1.4845	0.2231

rameterizations, Type 3 effect tests are replaced by joint tests. The joint test for an effect is a test that all of the parameters associated with that effect are zero equivalent to Type 3 effect tests under GLM parameterization.

Analysis of Maximum Likelihood Estimates

Parameter		DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits		Label
treat	0	1	-0.34824	0.49279	0.4994	0.4798	.	.	.	treat 0
age		1	-0.02272	0.01062	4.5774	0.0324	.	.	.	
age*treat	0	1	0.01832	0.01504	1.4845	0.2231	.	.	.	treat 0 * age

The following table summarizes the results of running a crude model: only treat variable, the main effects model: treat with covariate age, and the interaction model with treat, age and age x treat.

Model	Variable	**Coefficient**	**p-value**
Crude	treat	0.23744	0.0107
Main Effects	treat age	0.24165 -0.01370	0.0094 0.0674
Interaction	treat age treat*age	-0.34824 -0.02272 0.01832	0.4798 0.0324 **0.2231**

The interaction of treatment and age are not significant (p-value = 0.2231) in the interaction model; therefore, age is not an effect modifier. Hence, the effect of treatment (both short and long) will be same for all ages. Additionally, checking the % change in coefficient:

$\widehat{\Delta\beta}$**%** = ((0.23744 - 0.24165)/ 0.24165) * 100% = -**1.74% is small**

→ Hence, age is not a confounder of the association of treatment with recurrence time to drug use.

Furthermore, as we saw above, the mean age of participants for treatment = Short is 32.66 years and the mean age of participants for treatment=Long is 32.10 years. Since the difference is small, age is not a confounder of treatment. However, since age is borderline not a significant (p-value=0.0674) covariate, we can interpret the adjusted hazard ratio from the main effects model. The estimated hazard ratio for treatment is $\mathbf{e^{0.24165} = 1.273}$. This means that participants who received short treatment are 27.3% more likely to experience a recurrence to drug use, after adjusting for age.

Now, we look at the treat variable with treatment site to see if site is an effect modifier or confounder.

<u>After fitting the Cox PH model with treat and site variable to model recurrence time we get the following output</u>:

Model Fit Statistics

Criterion	Without Covariates	With Covariates
-2 LOG L	4978.748	4970.804
AIC	4978.748	4974.804
SBC	4978.748	4983.084

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	7.9435	2	0.0188
Score	7.9588	2	0.0187
Wald	7.9276	2	0.0190

Type 3 Tests

Effect	DF	Wald Chi-Square	Pr > ChiSq
treat	1	6.8317	0.0090
sit	1	1.4140	0.2344

Analysis of Maximum Likelihood Estimates

Parameter		DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits		Label
treat	0	1	0.24368	0.09323	6.8317	0.0090	1.276	1.063	1.532	treat 0
sit	0	1	0.12110	0.10184	1.4140	0.2344	1.129	0.924	1.378	sit 0

Looking at the p-values in the table above, we can see that the treat is significant, but the covariate site is not significant at 0.05 level.

Similarly, as above, I went ahead to test if site was a confounder or an effect modifier in the recurrence model. To do so I also need to fit an interaction model of site x treat and the following is the output:

Model Fit Statistics

Criterion	Without Covariates	With Covariates
-2 LOG L	4978.748	4969.905
AIC	4978.748	4975.905
SBC	4978.748	4988.325

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	8.8425	3	0.0315
Score	9.1090	3	0.0279
Wald	9.0478	3	0.0287

Joint Tests

Effect	DF	Wald Chi-Square	Pr > ChiSq
treat	1	0.4013	0.5264
sit	1	0.0171	0.8960
treat*sit	1	0.9005	0.3426

Note: Under full-rank parameterizations, Type 3 effect tests are replaced by joint tests. The joint test for an effect is a test that all of the parameters associated with that effect are zero. Such joint tests might not be equivalent to Type 3 effect tests under GLM parameterization.

Analysis of Maximum Likelihood Estimates

Parameter			DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits		Label
treat	0		1	0.10800	0.17049	0.4013	0.5264	.	.	.	treat 0
sit	0		1	0.01916	0.14649	0.0171	0.8960	.	.	.	sit 0
treat*sit	0	0	1	0.19300	0.20338	0.9005	0.3426	.	.	.	treat 0 * sit 0

The below table summarizes the results of running a crude model: only treat variable, the main effects model: treat with covariate site, and the interaction model with treat, site and site x treat.

Model	Variable	Coefficient	p-value
Crude	treat	0.23744	0.0107
Main Effects	treat site	0.24368 0.12110	0.0090 0.2344
Interaction	treat site treat*site	0.10800 0.01916 0.19300	0.5264 0.8960 **0.3426**

The interaction of treatment and site are not significant (p-value = 0.3426) in the interaction model; therefore, size is not an effect modifier. The effect of treatment (both short and long) will be same for all treatment sites.
Checking the % change in coefficient:

$\widehat{\Delta\beta}$**% = ((0.23744 - 0.24368)/ 0.24368) * 100% = -2.56% is small**

→ Hence, site is not a confounder of the association of treatment with recurrence time to drug use.

Since the difference is small, site is not a confounder of treatment. However, since site is not a significant (p-value=0.2344) covariate, we can interpret the adjusted hazard ratio from the main effects model. The estimated hazard ratio for treatment is $e^{0.24368} = 1.276$.

Similarly, this means that participants who received short treatment are 27.6% more likely to experience a recurrence to drug use, after adjusting for treatment site.

---

iii.

Building the best PH model for recurrence to drug use:

After testing out few variables, i.e., age and site are neither effect modifiers or confounder of treatment in relation to time to recurrence, I wanted to go ahead and build a best Proportional Hazard model for modeling the recurrence time to drug use.

For developing our best PH model for modeling the recurrence time to drug use, I started with **stepwise selection** procedure where I fit the model with all main effects and set the significance level for entry and exit out of the model to be 0.2.

- The following is the summary of the stepwise process:

Summary of Stepwise Selection

Step	Effect Entered	Effect Removed	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq	Effect Label
1	los		1	1	105.1038		<.0001	los
2	race		1	2	15.9851		<.0001	race
3	ndrugtx		1	3	13.4933		0.0002	
4	sit		1	4	10.8180		0.0010	sit
5	ivhx		2	5	9.8782		0.0072	ivhx
6	age		1	6	6.8802		0.0087	
7	treat		1	7	2.3623		0.1243	treat

The main effects fit into the model were *treat, race, age, ivhx, ndrugtx, site, los, beck and hercoc.* After running stepwise selection, *beck* and *hercoc* were discarded as they were the least important variables in the model for recurrence time to drug use.

Now, running the model with only the selected variables,

- The following is the output of the selected main effects model form the stepwise procedure:

Analysis of Maximum Likelihood Estimates									
Parameter		DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits	Label
treat	0	1	-0.14803	0.09604	2.3759	0.1232	0.862	0.714 1.041	treat 0
race	0	1	0.31259	0.11491	7.4000	0.0065	1.367	1.091 1.712	race 0
age		1	-0.02086	0.00817	6.5131	0.0107	0.979	0.964 0.995	
ivhx	1	1	-0.46536	0.12060	14.8906	0.0001	0.628	0.496 0.795	ivhx 1
ivhx	2	1	-0.26443	0.13283	3.9628	0.0465	0.768	0.592 0.996	ivhx 2
ndrugtx		1	0.02608	0.00849	9.4421	0.0021	1.026	1.009 1.044	
sit		1	0.43530	0.11043	15.5384	<.0001	1.545	1.245 1.919	sit
los		1	-0.00939	0.0008131	133.2764	<.0001	0.991	0.989 0.992	los

- Looking at the model selected from stepwise above, the *treat* variable is not significant (p-value = 0.1232). I also wanted to check for potential interaction effects hence I will leave the *treat* variable in the model for now.

Moreover, I wanted to test the Treatment Site(*sit*) against each other covariates to identify if there was an interaction in locations and other covariates when modeling for recurrence time to drug use. Along with this, I also checked for any interaction between Length of Treatments(*los*) and each other covariates as it may be of clinical interest.

The significance level for entry and exit for interaction terms out of the model is set to 0.1.

- <u>The following is the summary of the stepwise selection of main effects along with interaction terms followed by the output of the estimates:</u>

Summary of Stepwise Selection								
Step	Effect Entered	Effect Removed	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq	Effect Label
1	los		1	1	105.1038		<.0001	los
2	race		1	2	15.9851		<.0001	race
3	ndrugtx		1	3	13.4933		0.0002	
4	sit		1	4	10.8180		0.0010	sit
5	ivhx		2	5	9.8782		0.0072	ivhx
6	age		1	6	6.8802		0.0087	
7	los*ivhx		2	7	9.7708		0.0076	
8	age*sit		1	8	5.8998		0.0151	
9	ndrugtx*sit		1	9	4.0709		0.0436	

- We can see that *treat* was removed from the model which makes sense under stricter significance entry and exit level. Also, out of the multiple interactions tested, tested 12 in total (sit with all other variables and los with all other variables), 3 were selected from stepwise. The interactions selected were *los*ivhx, age* sit, ndrugtx*sit*.

<u>The following is the Maximum Likelihood estimates summary table:</u>

Parameter		DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq
race	0	1	0.31848	0.11570	7.5767	0.0059
age		1	-0.03892	0.01009	14.8890	0.0001
ivhx	1	1	-0.71374	0.20124	12.5793	0.0004
ivhx	2	1	-0.78141	0.22211	12.3774	0.0004
ndrugtx		1	0.03890	0.00971	16.0455	<.0001
sit		1	-0.86217	0.50996	2.8583	0.0909
los		1	-0.01193	0.00131	82.5044	<.0001
los*ivhx	1	1	0.00288	0.00183	2.4767	0.1155
los*ivhx	2	1	0.00575	0.00185	9.6569	0.0019
age*sit		1	0.04532	0.01592	8.1001	0.0044
ndrugtx*sit		1	-0.03926	0.01956	4.0273	0.0448

- From the above resulting output, now, I will start removing interaction terms one by one which have a p-value of greater than the 0.05 significance level until all the parameters in the model are significant to find the final model.

- Since, *los*ivhx* interaction has the highest insignificant p-value (0.1155) we will remove this interaction from the model and fit the proportional hazard model again. We get:

Parameter		DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq
race	0	1	0.32059	0.11529	7.7324	0.0054
age		1	-0.03942	0.01016	15.0541	0.0001
ivhx	1	1	-0.45178	0.12099	13.9427	0.0002
ivhx	2	1	-0.24651	0.13385	3.3919	0.0655
ndrugtx		1	0.03804	0.00961	15.6708	<.0001
sit		1	-0.92362	0.50945	3.2869	0.0698
los		1	-0.00930	0.0008043	133.6677	<.0001
age*sit		1	0.04703	0.01591	8.7418	0.0031
ndrugtx*sit		1	-0.03543	0.01940	3.3345	0.0678

- Since *ndrugtx*sit* interaction has a high insignificant p-value (0.0678), we will remove this interaction from the model and fit the proportional hazard model again.

Parameter		DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq
race	0	1	0.33210	0.11527	8.3012	0.0040
age		1	-0.03603	0.01000	12.9957	0.0003
ivhx	1	1	-0.43105	0.11968	12.9711	0.0003
ivhx	2	1	-0.26182	0.13279	3.8875	0.0486
ndrugtx		1	0.02820	0.00854	10.9135	0.0010
sit		1	-0.86162	0.50758	2.8816	0.0896
los		1	-0.00916	0.0007970	132.1763	<.0001
age*sit		1	0.03985	0.01537	6.7191	0.0095

- Preserving model hierarchy, although *sit* has the highest p-value, I will remove, age*sit interaction and fit the proportional hazard model again.

- Besides, IVHX which is borderline significant, all the parameters have p-values less than 0.05. Additionally, no significant interaction terms made it into the model. Hence, the following is the final model for recurrence time to drug use:

Analysis of Maximum Likelihood Estimates										
Parameter		DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits		Label
race	0	1	0.29251	0.11434	6.5443	0.0105	1.340	1.071	1.676	race 0
age		1	-0.02151	0.00818	6.9120	0.0086	0.979	0.963	0.995	
ivhx	1	1	-0.44998	0.12023	14.0065	0.0002	0.638	0.504	0.807	ivhx 1
ivhx	2	1	-0.25565	0.13300	3.6948	0.0546	0.774	0.597	1.005	ivhx 2
ndrugtx		1	0.02683	0.00849	9.9901	0.0016	1.027	1.010	1.044	
sit	0	1	-0.41190	0.11018	13.9744	0.0002	0.662	0.534	0.822	sit 0
los		1	-0.00915	0.0008083	128.1570	<.0001	0.991	0.989	0.992	los

The final cox proportional hazards model is:

$$\mathbf{h\ (t, x, \beta)} = h_0\ (t) * \exp\ (\beta_{Race=0} * x_1 + \beta_{Age} * x_2 + \beta_{IVHX=1} * x_3 + \beta_{IVHX=2} * x_4 + \beta_{Ndrugtx} * x_5 + \beta_{sit=0} * x_6 + \beta_{los} * x_7)$$

where,

h (t, x, β) represents hazard at time t
h_0 (t) represents baseline hazard function when all the predictors equal 0.
X_i, represents covariates of race, age, IVHX=1(never), IVHX=2(previous), ndrugtx, sit and los respectively, where i = 1,2,3,4,6,7
βi, represents parameter coefficients of the covariates

- Before proceeding with the final model, it is necessary to check the model assumptions.

4. MODEL CHECKING

- **CHECKING MODEL ADEQUACY**

We need to check the PH assumption.

- That is, the HR for each covariate in the model is independent of time. If the covariates vary with time, then the proportional hazard assumption is violated. Hence, I will test if the covariates are time dependent.

a) We can test the PH assumption by include time x covariate interaction terms one at a time and evaluate the significance. If any of these interaction terms are significant then the PH assumption is violated.

Five models were run to test time dependence:

Model 1: Race, Age, IVHX, NdrugTX, Site, Los, **Age*time**
Model 2: Race, Age, IVHX, NdrugTX, Site, Los, **IVHX*time**
Model 3: Race, Age, IVHX, NdrugTX, Site, Los, **NdrugTX*time**
Model 4: Race, Age, IVHX, NdrugTX, Site, Los, **Site*time**
Model 5: Race, Age, IVHX, NdrugTX, Site, Los, **Los*time**

- I will not test the **race** variable because it does not vary with time.

The results for testing interaction with each covariates one at a time in the final model are:

Interaction with time	interaction chi-square p-value
Age*time	0.6008
IVHX*time	0.3858
Ndrugtx*time	0.1989
Site*time	0.1167
Los*time	<0.0001

From the table above, only the interaction of **los x time** is significant. All the other variables are not significant, hence, the PH assumption for only los, length of treatment, seems to be violated. However, it does make sense that length of treatment varies with time, as the number of days a participant is treated over time is different.

- Let us look further into the PH assumption by plotting survival curves using levels or strata for the covariates in our model, especially focusing on the continuous covariates.

b) Looking at PH assumptions graphically though survival curves.

Starting off with the ***continuous*** covariates, we need to create stratums to plot survival curves for each level. I will use quartiles from below for creating these levels.

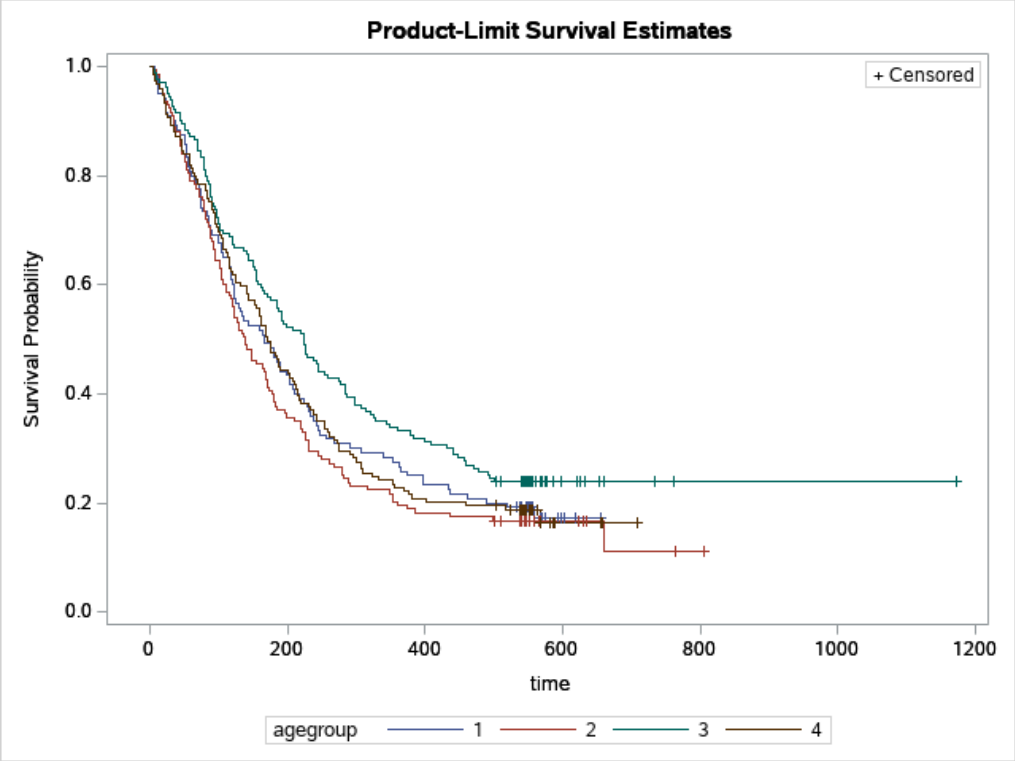
The MEANS Procedure

Variable	Label	Lower Quartile	Median	Upper Quartile
age		27.0000000	32.0000000	37.0000000
ndrugtx		1.0000000	3.0000000	6.0000000
los	los	45.0000000	87.0000000	151.0000000

1) KM curve for **agegroup**:

Age was divided into 4 strata:

- Age < 27, agegroup =1
- 27 <= Age < 32, agegroup =2
- 32 <= Age < 37, agegroup =3
- 37 <= Age, agegroup =4

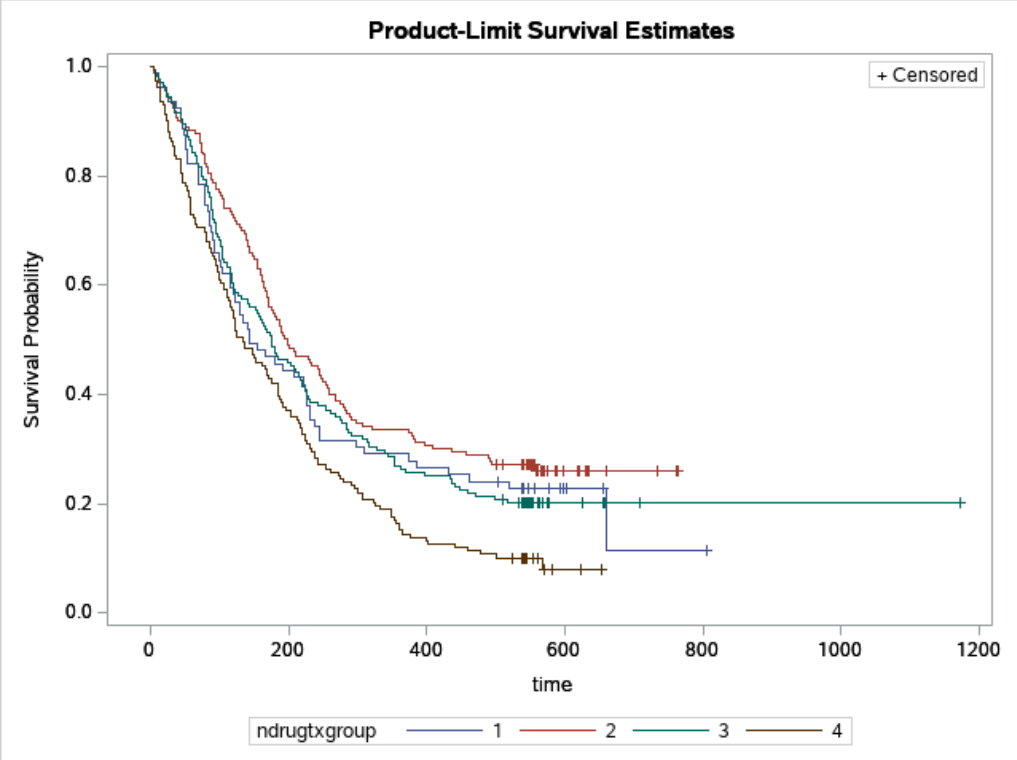


From the graph, we can see that survival curve for age group 3 is the highest. The curves do follow a similar trend for most of the part, but a point of concern is when the curve for age group 1 crosses the other curves which could potentially violate the PH assumption.

2) KM curve for **ndrugtx group**:

Ndrugtx was divided into 4 strata:

- Ndrugtx < 1, ndrugtxgroup =1
- 1 <= Ndrugtx < 3, ndrugtxgroup =2
- 3 <= Ndrugtx < 6, ndrugtxgroup =3
- 6 <= Ndrugtx, ndrugtxgroup =4

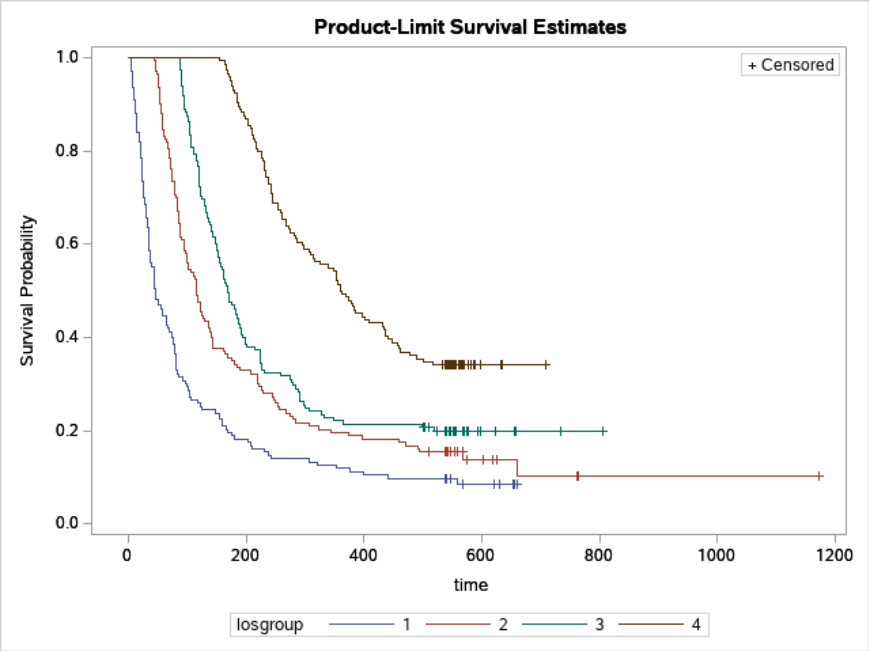


From the graph, we can see that survival curve for number of prior drug treatments group 2 is the highest. The curves do follow a similar trend for most of the part, but again a point of concern is when the curves cross paths which might potentially violate the PH assumption.

3) KM curve for **los group**:

Los was divided into 4 strata:

- Los < 45, losgroup =1
- 45 <= Los < 87, losgroup =2
- 87 <= Los < 87, losgroup =3
- 151 <= Los, losgroup =4

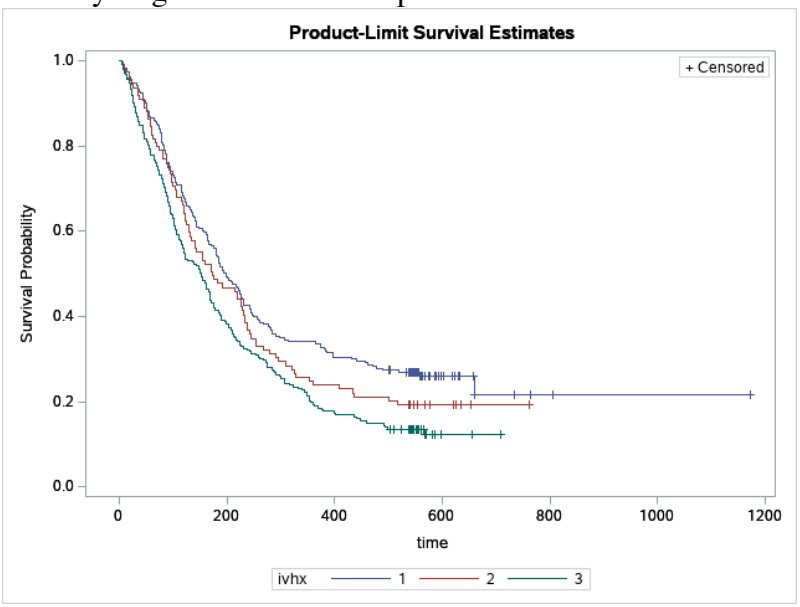


From the graph above, that survival curve for length of treatment group 4 is the highest. The curve for group 4 diverges a lot quicker compared to other groups which might potentially violate the PH assumption.

Now, let's look at the ***categorical*** covariates k-m curves:

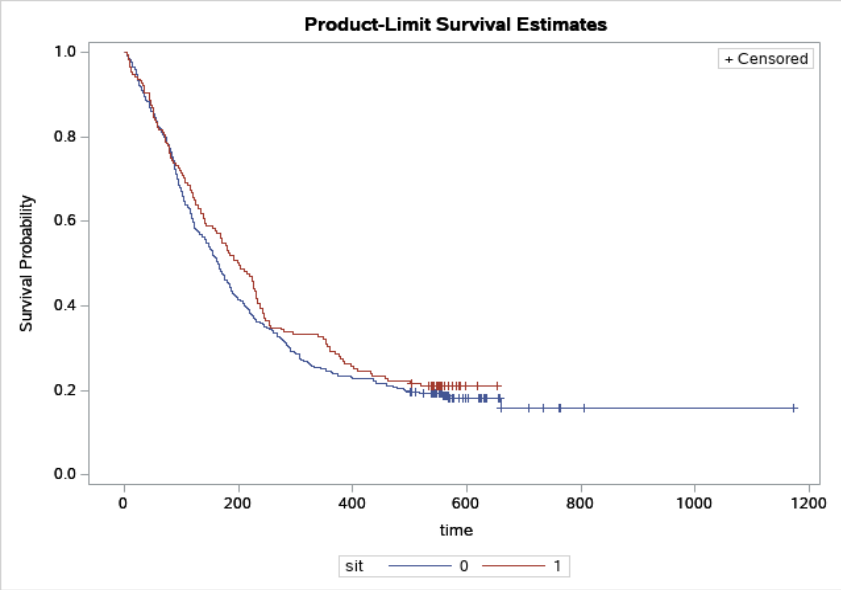
4) KM curve for **IVHX**

The survival curve for IVHX= 1(recent use history) is the highest. The curves have the same basic shape and start closely together, but IVHX=2 category seems to cross category 1 at the early stages. The PH assumption seems to be somewhat violated here.



5) KM curve **for Site**

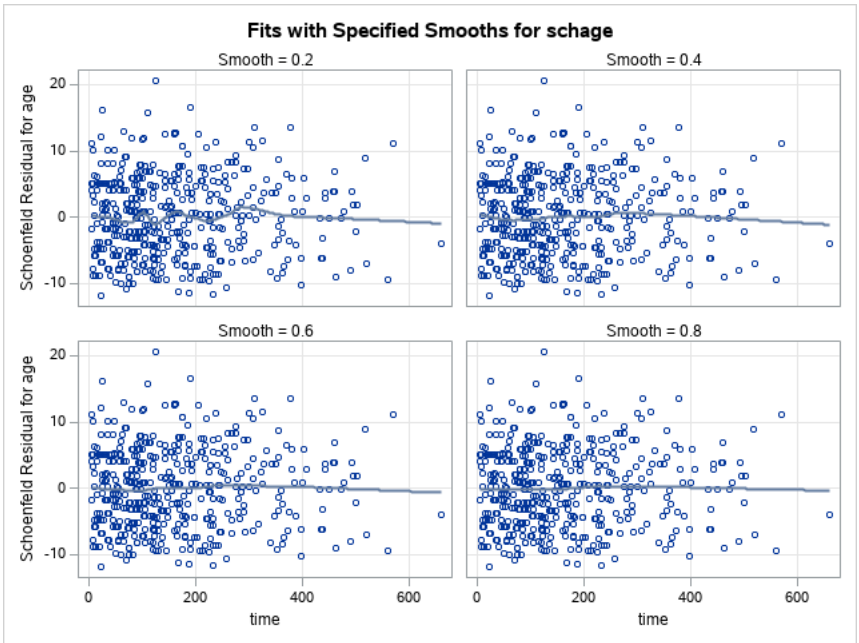
The survival curve for site= 1 is the highest. The curves have the same basic shape and start closely together, but site 1 comes closer to site 0 towards the middle. The PH assumption seems to be somewhat concerning here.



c) Check PH assumptions for covariates by plotting **Schoenfeld residuals vs. Functions of time.**

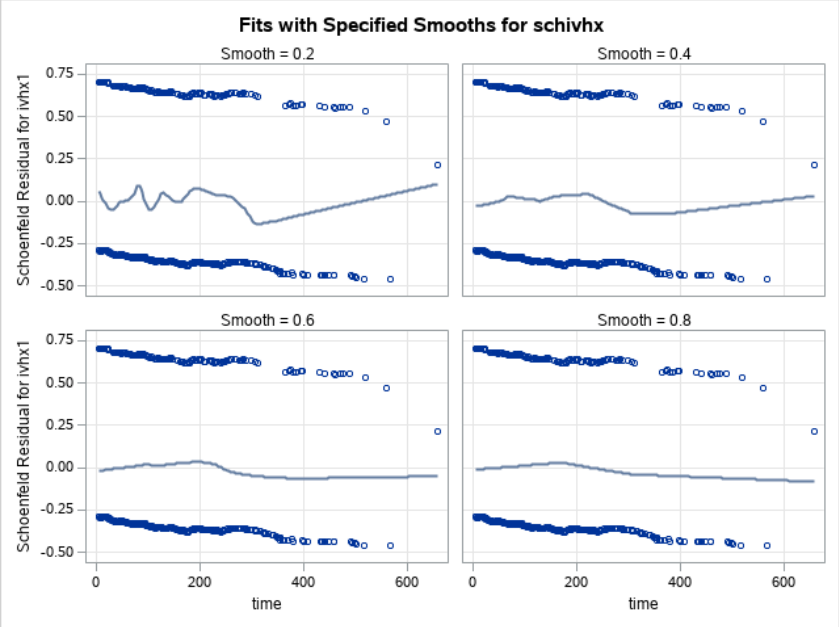
Age

Towards the beginning at smooth=0.2, it's not proportional. As we increase the smooth the expected value comes towards 0.



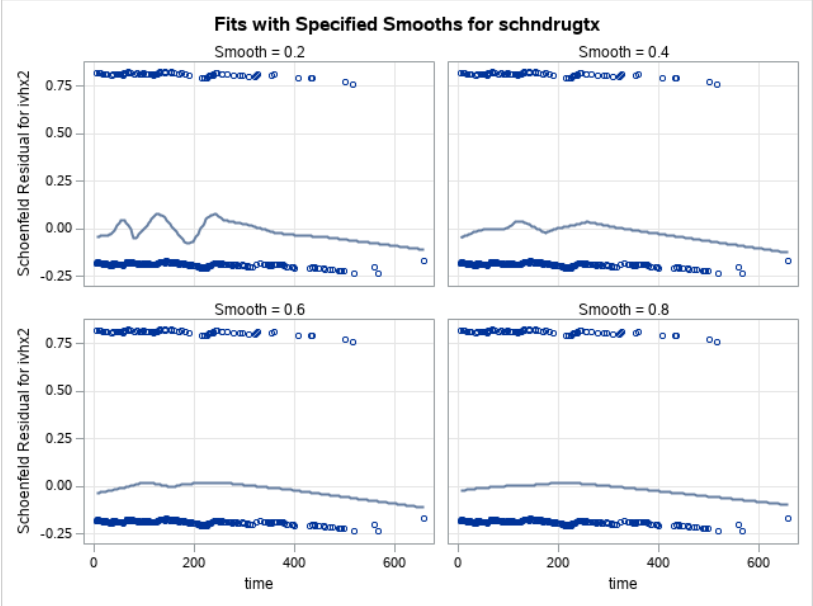
IVHX

The loess smoother is bumpy towards the beginning, the expected value is less than 0; not proportional.



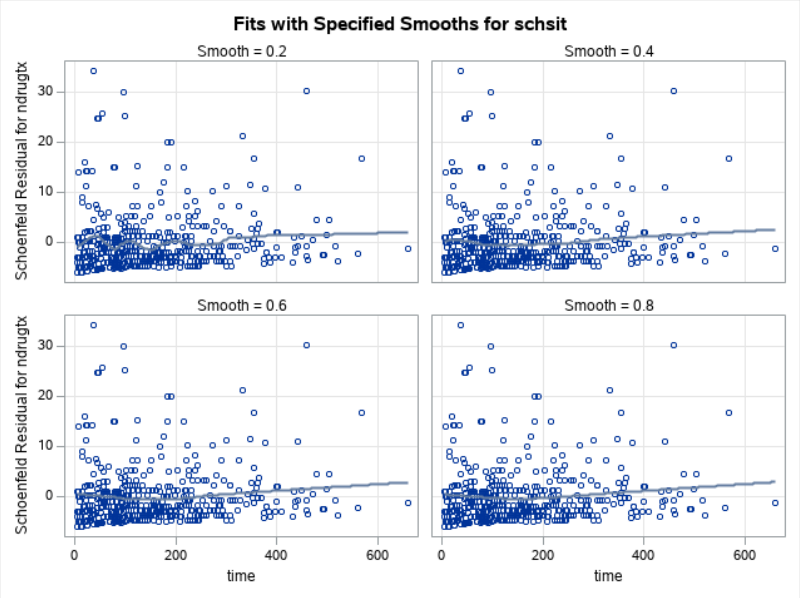
NDRUGTX

Similar to above, the loess smoother is bumpy towards the beginning, the expected value is less than 0; not proportional.



Site

The expected value over time is more than 0, even at smooth =0.8; not proportional.

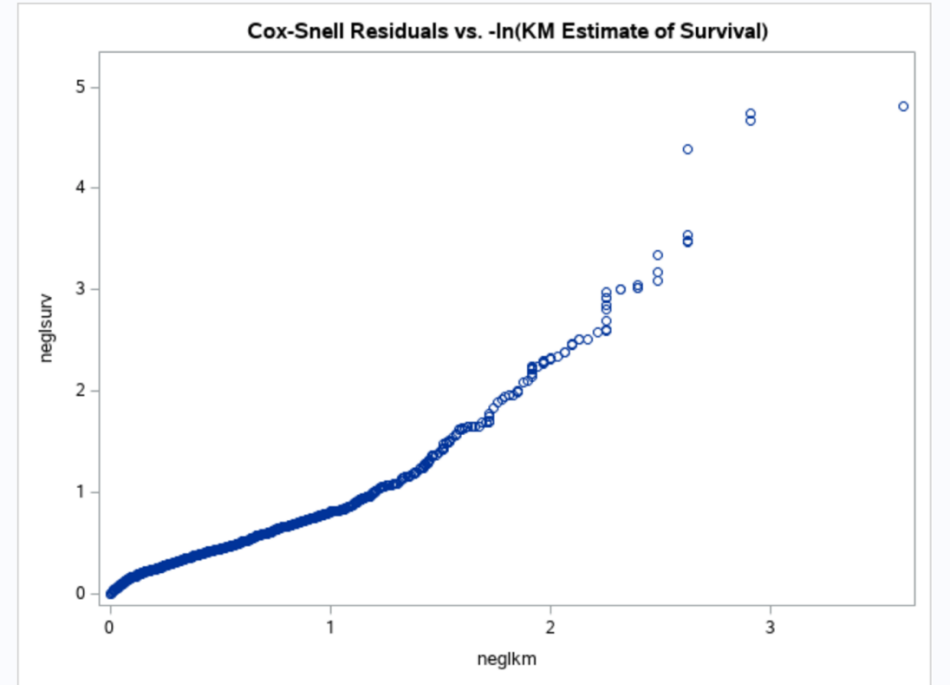


Los

The loess smoother is bumpy towards the beginning, the expected value is less than 0; not proportional.

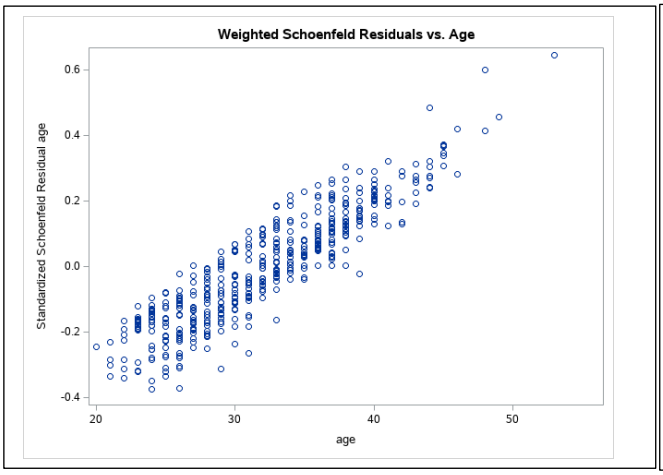
- **ASSESSING GOODNESS OF FIT BY RESIDUALS**

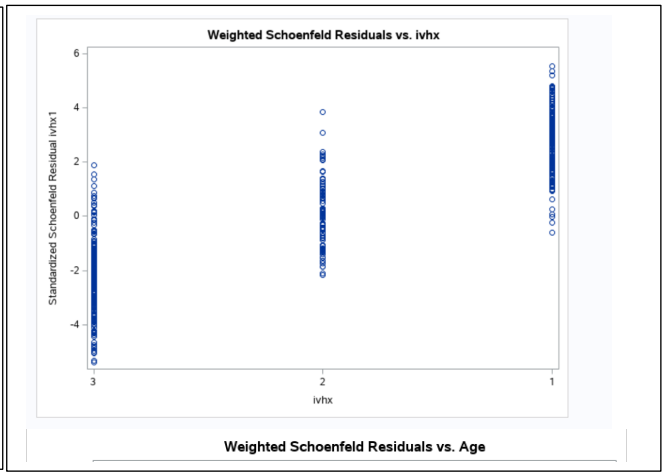
  - **Cox-Snell Residuals**

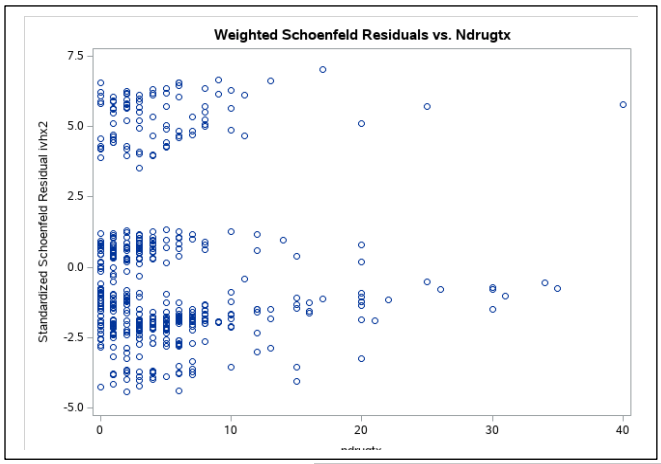


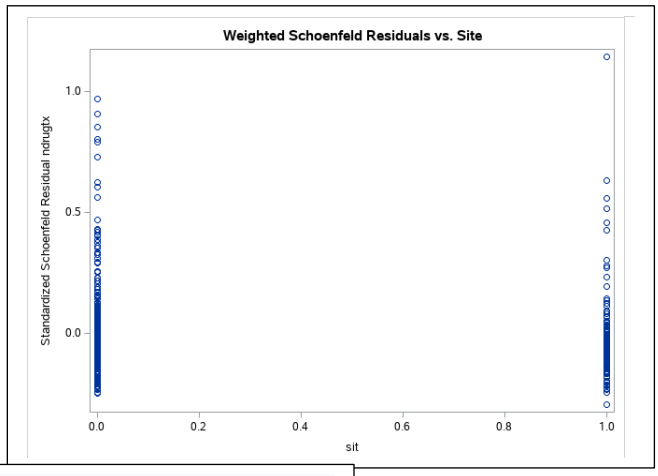
➔ Generally, follows a 45-degree line until 1.5 *neglkm*, then falls above the line. However, only few residuals past 1.5, the assumption for the goodness of fit seems to have been met.

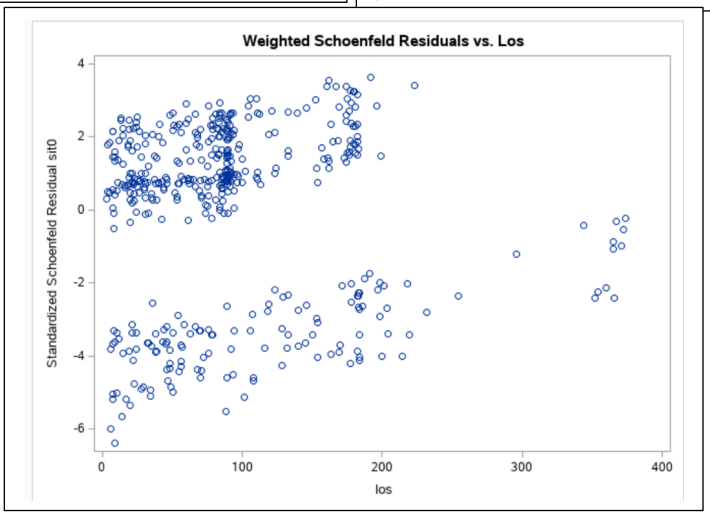
- **Schoenfeld Residuals**











➔ Additionally. looking at all the weighted Schoenfeld residuals below, all the residual plots seem to be symmetric about 0. No clear patterns seem to exist except for some time trends as expected. Hence, we arrive at the same conclusion as the Cox-Snell Residuals. The assumption of goodness of fit by residuals is met.

5. CONCLUSIONS

When modeling recurrence time to drug use with treatment type, the long treatment seemed more statistically significant and effective than the short treatment assignment.

Additionally, age of the participants and the treatment site were neither confounders nor effect modifiers when paired with treatment type for modeling the recurrence time to drug use.

Furthermore, the best model for modeling the recurrence to drug use has race, age, IV drug use history, number of prior drug treatments, treatment site and length of treatment as covariates.
After checking for model assumptions, although the residuals test gives an indication that the assumption of goodness of fit by residuals is appropriate, the PH assumptions for almost all of the covariates when checking graphically (Schoenfeld residuals and KM curves) seems to be somewhat violated. In the case of checking the PH assumption through interactions of each variable (one at a time) with time, the length of treatment seems to be the only one violated as it had a high p-value.

As a result, for future work it is better to use a proportional hazards model with time dependent covariates for this dataset. Since that is beyond our project requirement, I will still use the final best model selected here to make inferences about the coefficients. However, please keep in mind that the coefficient estimated is a sort of average effect over time.

Analysis of Maximum Likelihood Estimates

Parameter		DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits		Label
race	0	1	0.29251	0.11434	6.5443	0.0105	1.340	1.071	1.676	race 0
age		1	-0.02151	0.00818	6.9120	0.0086	0.979	0.963	0.995	
ivhx	1	1	-0.44998	0.12023	14.0065	0.0002	0.638	0.504	0.807	ivhx 1
ivhx	2	1	-0.25565	0.13300	3.6948	0.0546	0.774	0.597	1.005	ivhx 2
ndrugtx		1	0.02683	0.00849	9.9901	0.0016	1.027	1.010	1.044	
sit	0	1	-0.41190	0.11018	13.9744	0.0002	0.662	0.534	0.822	sit 0
los		1	-0.00915	0.0008083	128.1570	<.0001	0.991	0.989	0.992	los

Looking at each variable in our model, the following are the interpretations for each of the hazard ratios:

Race

- White individuals have 34% higher chance of returning to drug use than other individuals after adjusting for other factors in the model.

Age

- For every 1-year increase in age of the participants, the risk of returning to drug use decreases by 2.1% after adjusting for other factors in the model.

IVHX

IVHX=1 (never)

- Participants who never had an IV drug use history have 36.2% lower chance of returning to drug use than participants who have had a recent IV drug use history after adjusting for other factors in the model.

IVHX=2 (previous)

- Participants who previously had an IV drug use history have 22.6% lower chance of returning to drug use than participants who have had a recent IV drug use history after adjusting for other factors in the model

Ndrugtx

- For each additional prior drug treatment, the risk of returning to drug use increases by 2.7% after adjusting for other factors in the model.

Site

- Participants at Site A have 33.8% lower chance of returning to drug use than participants at site B after adjusting for other factors in the model.

Los

- For every 1-day increase in the length of the treatment of the participants, the risk of returning to drug use decreases by 0.9% after adjusting for other factors in the model.

6. APPENDIX

a. SAS code

```
/*************************************************************************/
MAT 8446 - SURVIVAL DATA ANALYSIS PROJECT
/*************************************************************************/

FILENAME REFFILE '/folders/myfolders/uis.xlsx';

PROC IMPORT DATAFILE=REFFILE REPLACE
        DBMS=XLSX
        OUT=UIS0;
        GETNAMES=YES;
RUN;

PROC PRINT DATA= UIS0 (OBS=10);

PROC CONTENTS DATA = UIS0;

/*FIX THE FORMAT FOR VARIABLES */

DATA UIS1;
 SET UIS0;
 age0= INPUT(age, 8.);
 beck0= INPUT(beck, 8.);
 ndrugtx0= INPUT(ndrugtx, 8.);
 DROP AGE BECK NDRUGTX;
 RENAME age0= age
beck0= beck ndrugtx0=ndrugtx;
RUN;

PROC PRINT DATA= UIS1 (OBS=10);
PROC CONTENTS DATA= UIS1;


/*CHECK FOR MISSING VALUES */
PROC FREQ DATA=UIS1;
TABLES HERCOC IVHX RACE;
RUN;

PROC MEANS DATA=UIS1 NMISS N;
RUN;

/*REMOVE ROWS WITH MISSING VALUES */
DATA UIS;
 SET UIS1;
 IF HERCOC = '' THEN DELETE;
 IF IVHX = '' THEN DELETE;
 IF RACE = . THEN DELETE;
 IF AGE = . THEN DELETE;
 IF BECK = . THEN DELETE;
 IF NDRUGTX = . THEN DELETE;

RUN;

/*CHECK IF MISSING VALUES REMOVED */
PROC FREQ DATA=UIS;
TABLES HERCOC IVHX RACE;
RUN;

PROC MEANS DATA=UIS NMISS N;
RUN;
*ok no missing values;

proc print data=uis (obs=50);
```

```
*1;

PROC LIFETEST DATA=UIS METHOD=KM PLOTS=S(CL ATRISK(ATRISKTICK)) TIMELIST = 365 730 1095 1460
REDUCEOUT OUTSURV=out1 CONFTYPE=LOGLOG;
  TITLE 'Comparison of Multiple Survival Functions';
  TIME TIME*CENSOR(0);
  STRATA TREAT / ADJUST = TUKEY ;
RUN;

PROC PRINT DATA=out1;
  TITLE '1-, 2-, 3-, and 4-Year Survival Estimates with CIs';
RUN;


*2;

PROC MEANS DATA= UIS;
CLASS TREAT;
VAR AGE;
RUN;


/* CRUDE MODEL*/
PROC PHREG DATA= UIS;
 TITLE 'MODEL WITH ONE CONTINUOUS COVARIATE- CRUDE';
 CLASS TREAT;
 MODEL TIME*CENSOR(0)=TREAT/ RL=WALD TIES=EXACT;
RUN;

/* MAIN EFFECTS MODEL*/
PROC PHREG DATA= UIS;
 TITLE 'MODEL WITH ONE CONTINUOUS COVARIATE-MAIN EFFECTS';
 CLASS TREAT;
 MODEL TIME*CENSOR(0)= TREAT AGE/ RL=WALD ties=exact ;
RUN;

/* INTERACTION MODEL*/
PROC PHREG DATA= UIS;
 TITLE 'MODEL WITH ONE CONTINUOUS COVARIATE- INTERACTION';
 CLASS TREAT;
 MODEL TIME*CENSOR(0)= TREAT AGE TREAT*AGE/ RL=WALD TIES=EXACT;
RUN;



*3;

*SITE-Categorical;

/* CRUDE MODEL*/
PROC PHREG DATA= UIS ;
 TITLE 'MODEL WITH ONE CATEGORICAL COVARIATE- CRUDE';
 CLASS TREAT;
 MODEL TIME*CENSOR(0)=TREAT/ RL=WALD TIES=EXACT;
RUN;

/* MAIN EFFECTS MODEL*/
PROC PHREG DATA= UIS;
 TITLE 'MODEL WITH ONE CATEGORICAL COVARIATE-MAIN EFFECTS';
 CLASS TREAT SIT;
 MODEL TIME*CENSOR(0)= TREAT SIT/ RL=WALD TIES=EXACT;
RUN;

/* INTERACTION MODEL*/
PROC PHREG DATA= UIS;
 TITLE 'MODEL WITH ONE CATEGORICAL COVARIATE- INTERACTION';
 CLASS TREAT SIT;
 MODEL TIME*CENSOR(0)= TREAT SIT TREAT*SIT/ RL=WALD TIES=EXACT;
```

```
RUN;


*4;

*The following runs the forward selection procedure;
PROC PHREG DATA=UIS;
 TITLE "Stepwise Selection";
 CLASS TREAT RACE HERCOC IVHX;
 MODEL TIME*CENSOR(0)= TREAT RACE AGE BECK HERCOC IVHX NDRUGTX SIT LOS / SELECTION=STEPWISE
SLENTRY=0.20 SLSTAY=0.20 DETAILS;
RUN;

PROC PHREG DATA=UIS;
 TITLE "Final Selected Main effects";
 CLASS TREAT RACE IVHX;
 MODEL TIME*CENSOR(0)= TREAT RACE AGE IVHX NDRUGTX SIT LOS/ RL=WALD TIES=EXACT;
RUN;

*Interaction forward selection procedure;
PROC PHREG DATA=UIS;
 TITLE "Stepwise Selection";
 CLASS TREAT RACE IVHX;
 MODEL TIME*CENSOR(0)= TREAT RACE AGE IVHX NDRUGTX SIT LOS
/*        AGE*TREAT AGE*RACE AGE*IVHX AGE*NDRUGTX AGE*SIT AGE*LOS */
/*        RACE*TREAT RACE*IVHX RACE*NDRUGTX RACE*SIT RACE*LOS */
       LOS*IVHX SIT*TREAT SIT*IVHX SIT*NDRUGTX SIT*LOS AGE*SIT RACE*SIT/ SELECTION=STEPWISE
SLENTRY=0.1 SLSTAY=0.1 DETAILS;
RUN;

PROC PHREG DATA=UIS;
 TITLE "Final Selected";
 CLASS TREAT RACE IVHX;
 MODEL TIME*CENSOR(0)= RACE AGE IVHX NDRUGTX SIT LOS
         LOS*IVHX AGE*SIT NDRUGTX*SIT/ RL=WALD TIES=EXACT;
RUN;
PROC PHREG DATA=UIS;
 TITLE "Final Selected-remove los*ivhx";
 CLASS TREAT RACE IVHX;
 MODEL TIME*CENSOR(0)= RACE AGE IVHX NDRUGTX SIT LOS
         AGE*SIT NDRUGTX*SIT/ RL=WALD TIES=EXACT;
RUN;

PROC PHREG DATA=UIS;
 TITLE "Final Selected-remove los*ivhx";
 CLASS TREAT RACE IVHX;
 MODEL TIME*CENSOR(0)= RACE AGE IVHX NDRUGTX SIT LOS
         AGE*SIT NDRUGTX*SIT/ RL=WALD TIES=EXACT;
RUN;


PROC PHREG DATA=UIS;
 TITLE "Final Selected-remove age*sit";
 CLASS TREAT RACE IVHX SIT;
 MODEL TIME*CENSOR(0)= RACE AGE IVHX NDRUGTX SIT LOS/ RL=WALD TIES=EXACT;
RUN;

*FINAL MODEL;
PROC PHREG DATA=UIS;
 TITLE "Final Selected-remove ivhx";
 CLASS RACE SIT IVHX;
 MODEL TIME*CENSOR(0)= RACE AGE IVHX NDRUGTX SIT LOS/ RL=WALD TIES=EXACT;
RUN;

*******************************

CHECK FOR MODEL ADEQUACY

*******************************;
```

```
PROC PHREG DATA=UIS;
 TITLE "with race-time";
 CLASS RACE SIT IVHX;
 MODEL TIME*CENSOR(0)= RACE AGE IVHX NDRUGTX SIT LOS z/ RL=WALD TIES=EXACT;
 z = time*race;
RUN;


PROC PHREG DATA=UIS;
 TITLE "with age-time";
 CLASS RACE SIT IVHX;
 MODEL TIME*CENSOR(0)= RACE AGE IVHX NDRUGTX SIT LOS z/ RL=WALD TIES=EXACT;
 z = time*age;
RUN;

PROC PHREG DATA=UIS;
 TITLE "with ivhx-time";
 CLASS RACE SIT IVHX;
 MODEL TIME*CENSOR(0)= RACE AGE IVHX NDRUGTX SIT LOS z/ RL=WALD TIES=EXACT;
 z = time*ivhx;
RUN;

PROC PHREG DATA=UIS;
 TITLE "with ndrugtx-time";
 CLASS RACE SIT IVHX;
 MODEL TIME*CENSOR(0)= RACE AGE IVHX NDRUGTX SIT LOS z/ RL=WALD TIES=EXACT;
 z = time*ndrugtx;
RUN;

PROC PHREG DATA=UIS;
 TITLE "with sit-time";
 CLASS RACE SIT IVHX;
 MODEL TIME*CENSOR(0)= RACE AGE IVHX NDRUGTX SIT LOS z/ RL=WALD TIES=EXACT;
 z = time*sit;
RUN;

PROC PHREG DATA=UIS;
 TITLE "with los-time";
 CLASS RACE SIT IVHX;
 MODEL TIME*CENSOR(0)= RACE AGE IVHX NDRUGTX SIT LOS z/ RL=WALD;
 z = time*los;
RUN;


**PH Graphs check;

proc means data=uis  q1 median q3 ;
 var age ndrugtx los;
run;
*create stra for continuous variable;
data UIS00;
 set UIS;
 if age ne . and age<27 then agegroup =1;
 if 27<=age<32 then agegroup =2;
 if 32<=age<37 then agegroup =3;
 if 37<=age then agegroup =4;
 if ndrugtx ne . and ndrugtx<1 then ndrugtxgroup =1;
 if 1<=ndrugtx<3 then ndrugtxgroup =2;
 if 3<=ndrugtx<6 then ndrugtxgroup =3;
 if 6<=ndrugtx then ndrugtxgroup =4;
 if los ne . and los<45 then losgroup =1;
 if 45<=los<87 then losgroup =2;
 if 87<=los<151 then losgroup =3;
 if 151<=los then losgroup =4;
run;
****PH graphs*****;
PROC LIFETEST DATA=uis00 PLOTS=(S) METHOD = KM;
TITLE "KM Curves by Age Group";
```

```
TIME TIME*CENSOR(0);
STRATA agegroup;
RUN;

PROC LIFETEST DATA=uis00 PLOTS=(S) METHOD = KM;
TITLE "KM Curves by Ndrugtx Group";
TIME TIME*CENSOR(0);
STRATA ndrugtxgroup;
RUN;

PROC LIFETEST DATA=uis00 PLOTS=(S) METHOD = KM;
TITLE "KM Curves by Losgroup";
TIME TIME*CENSOR(0);
STRATA losgroup;
RUN;
*no need to check race;
PROC LIFETEST DATA=uis00 PLOTS=(S) METHOD = KM;
TITLE "KM Curves by race";
TIME TIME*CENSOR(0);
STRATA race;
RUN;

PROC LIFETEST DATA=uis00 PLOTS=(S) METHOD = KM;
TITLE "KM Curves by IVHX";
TIME TIME*CENSOR(0);
STRATA ivhx;
RUN;

PROC LIFETEST DATA=uis00 PLOTS=(S) METHOD = KM;
TITLE "KM Curves by site";
TIME TIME*CENSOR(0);
STRATA sit;
RUN;

**Schoenfeld Residuals**;
*FINAL MODEL;
PROC PHREG DATA=uis;
 TITLE "Final Model";
 CLASS RACE SIT IVHX;
 MODEL TIME*CENSOR(0)= RACE AGE IVHX NDRUGTX SIT LOS/ TIES=EXACT;
OUTPUT OUT=schoen RESSCH=schrace schage schivhx schndrugtx schsit scchlos;
RUN;

PROC LOESS DATA = schoen;
TITLE "Schoenfeld Residuals for Age vs. Time";
MODEL schage=time / SMOOTH=(0.2 0.4 0.6 0.8);
RUN;
PROC LOESS DATA = schoen;
TITLE "Schoenfeld Residuals for IVHX vs. Time";
MODEL schivhx=time / SMOOTH=(0.2 0.4 0.6 0.8);
RUN;
PROC LOESS DATA = schoen;
TITLE "Schoenfeld Residuals for Ndrugtx vs. Time";
MODEL schndrugtx=time / SMOOTH=(0.2 0.4 0.6 0.8);
RUN;
PROC LOESS DATA = schoen;
TITLE "Schoenfeld Residuals for Site vs. Time";
MODEL schsit=time / SMOOTH=(0.2 0.4 0.6 0.8);
RUN;
PROC LOESS DATA = schoen;
TITLE "Schoenfeld Residuals for Los vs. Time";
MODEL scchlos=time / SMOOTH=(0.2 0.4 0.6 0.8);
RUN;
*########################
*Residual check;
*########################;
PROC PHREG DATA=UIS;
 TITLE 'Proportional Hazards Model';
 CLASS RACE SIT IVHX;
```

```
 MODEL TIME*CENSOR(0)= RACE AGE IVHX NDRUGTX SIT LOS / RL=WALD;
 OUTPUT OUT=resids LOGSURV=lsurv RESDEV=deviance WTRESSCH= rrace rage rivhx rdrugtx rsit rlos;
RUN;

/* The following gets the negative of the log of the survival estimate for the Extended Cox-Snell
Residual*/
DATA resids;
SET resids;
neglsurv = -1*lsurv;
RUN;

/* The following gets the negative ln of the survival estimates based on Kaplan-Meier */
PROC LIFETEST DATA=resids OUTS=out1 NOPRINT;
TIME neglsurv*censor(0);
RUN;

DATA out1;
SET out1;
neglkm = -1*log(survival);
RUN;

/* The following creates the residual plot for the Extended Cox-Snell Residuals */
PROC SGPLOT DATA=out1;
TITLE 'Cox-Snell Residuals vs. -ln(KM Estimate of Survival)';
SCATTER Y= neglsurv X=neglkm;
RUN;

/* The following creates the Deviance Residual Plot */
/* PROC SGPLOT DATA=resids; */
/* TITLE 'Deviance Residuals vs. Time'; */
/* SCA deviance*survtime; */
/* PLOT deviance*age; */
/* RUN; */

/* The following creates the Weighted Schoenfeld Residual Plot */

%macro time(label= ,Y= ,X= );
PROC SGPLOT DATA=resids;
TITLE "Weighted Schoenfeld Residuals vs. &label";
SCATTER Y=&Y X=&X;
RUN;
%mend;

%time(label=Age,Y=rage , X=age );
%time(label=Race,Y=rrace , X=race );
%time(label=ivhx,Y=rivhx , X=ivhx );
%time(label=Ndrugtx,Y=Rdrugtx , X=Ndrugtx );
%time(label=Site,Y=rsit , X=sit );
%time(label=Los,Y=rlos , X=los );


*******************************
         TESTING-FINAL MODEL
*******************************;
*Testing individuals;
*****************************************;
PROC PHREG DATA= UIS;
 TITLE 'TREAT';
 CLASS TREAT ;
 MODEL TIME*CENSOR(0)= TREAT/ RL=WALD TIES=EXACT;
RUN;

PROC PHREG DATA= UIS;
 TITLE 'RACE';
 CLASS RACE ;
 MODEL TIME*CENSOR(0)= RACE/ RL=WALD TIES=EXACT;
RUN;

PROC PHREG DATA= UIS;
```

```
 TITLE 'HERCOC';
 CLASS HERCOC ;
 MODEL TIME*CENSOR(0)= HERCOC/ RL=WALD TIES=EXACT;
RUN;

PROC PHREG DATA= UIS;
 TITLE 'IVHX';
 CLASS IVHX ;
 MODEL TIME*CENSOR(0)= IVHX/ RL=WALD TIES=EXACT;
RUN;

PROC PHREG DATA= UIS;
 TITLE 'AGE';
 MODEL TIME*CENSOR(0)= AGE/ RL=WALD TIES=EXACT;
RUN;

PROC PHREG DATA= UIS;
 TITLE 'BECK';
 MODEL TIME*CENSOR(0)= BECK/ RL=WALD TIES=EXACT;
RUN;

PROC PHREG DATA= UIS;
 TITLE 'NDRUGTX';
 MODEL TIME*CENSOR(0)= NDRUGTX/ RL=WALD TIES=EXACT;
RUN;

PROC PHREG DATA= UIS;
 TITLE 'SITE';
 MODEL TIME*CENSOR(0)= SIT/ RL=WALD TIES=EXACT;
RUN;

PROC PHREG DATA= UIS;
 TITLE 'LENGTH OF TX';
 MODEL TIME*CENSOR(0)= LOS/ RL=WALD TIES=EXACT;
RUN;

********************* END ************************;
```